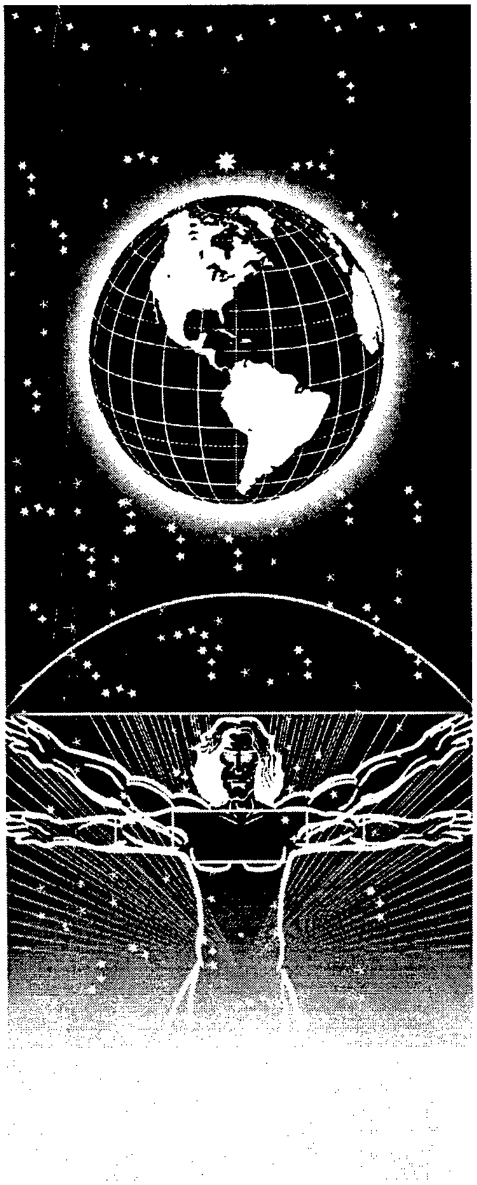
AL/HR-TR-1997-0131



UNITED STATES AIR FORCE RESEARCH LABORATORY

ASSESSING THE STABILITY OF STRUCTURAL LEARNING MEASURES

Martin Tessmer
University of South Alabama
Department of Behavioral Studies and
Educational Technology

Bruce Perrin
McDonnell-Douglas Aerospace
St. Louis, MO 63166-0566

Winston Bennett, Jr.

HUMAN RESOURCES DIRECTORATE
COGNITION AND PERFORMANCE DIVISION
7909 Lindbergh Drive
Brooks AFB, TX 78235-5352

October 1997

AIR FORCE MATERIEL COMMAND
AIR FORCE RESEARCH LABORATORY
HUMAN EFFECTIVENESS DIRECTORATE
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352

NOTICES

This report has been reviewed and is approved for publication.

WINSTON BENNETT, JR
Project Scientist

R. BRUCE GOULD
Technical Advisor

WILLIAM E. ALLEY
Chief
Cognition and Performance Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE September 1997	3. REPORT TYPE AND DATES COVERED Interim Report - September 1995-December 1996

4. TITLE AND SUBTITLE

Assessing the Stability of Structural Learning Measures

5. FUNDING NUMBERS

C-	F41622-96-P-2351
PE-	62202F
PR-	2300
TA-	HT
WU	61

6. AUTHOR(S)
Martin Tessmer
Bruce Perrin
Winston Bennett

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of South Alabama
Department of Behavioral Studies
and Educational Technology
Mobile, AL 36688

McDonnell Douglas Aerospace
St. Louis, MO 63166-0566

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Mission Critical Skills Division
7909 Lindbergh Drive
Brooks AFB, TX 78235-5352

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AL/HR-TR-1997-0131

11. SUPPLEMENTARY NOTES

Air Force Research Laboratory Technical Monitor: Winston Bennett (480) 988-6561, DSN 474-6297

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The study explored two related research questions: (1) Are structural learning measures stable? Sorting concepts printed on cards has traditionally been used for a variety of cognitive modeling applications; Pathfinder is a more recently developed developed, but widely used measure of mental model and concept network learning. Are these measures stable, based on test-retest indices? (2) As students learn more about the topic, does the structural learning measures' stability increase? Does the card sort or Pathfinder learning measure show increased stability as students' long-term memory structures develop?

14. SUBJECT TERMS
Education and Training Assessment
Knowledge Organization
Learning Measures
Training Effectiveness
Training Evaluation

15. NUMBER OF PAGES
28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION ABSTRACT
UNCLASSIFED	UNCLASSIFIED	UNCLASSIFIED	UL

CONTENTS

Figures

Tables

PREFACE

This research was supported under the Entrepreneurial Research program, Contract #F4162296P2351 from the Air Force, Human Effectiveness Directorate, Mission Critical Skills Division, Brooks AFB, TX. The opinions expressed in this paper are those of the authors and do not necessarily reflect the official position of the US Air Force. We would like to thank Mr. Mark Bowler for his review, comments, and revisions on earlier drafts of this report.

ASSESSING THE STABILITY OF STRUCTURAL LEARNING MEASURES

INTRODUCTION

Over the last several decades, cognitive psychology and instructional design have increasingly acknowledged the importance of structural learning. The learning process is described as the modification of memory structures (Anderson, 1995; Rumelhart & Norman, 1981). Structural learning outcomes have arisen such as mental models, conceptual networks, and information networks (Kraiger, Ford & Salas, 1993; Jonassen & Tessmer, in press). Researchers have investigated structural learning methods such as pattern noting, semantic mapping, and conceptual integration techniques (Jonassen, Beissner & Yacci, 1993).

With this growing importance on structural learning comes a growing focus upon structural learning measures. A structural learning measure identifies: a) the types of a concepts or propositions a learner has stored in long term memory; and b) the relationships between these concepts. These measures purportedly depict a learner's mental model or schema of complex topics, such as a mental model of a carburetor or a judicial system.

Structural measures of trainee knowledge are important because they reflect the difference between expert and novice knowledge structures (Shavelson, 1972). The degree of novice-expert difference on structural measures such as the groups of concepts formed in card sorting or Pathfinder networks has been predictive of differences in novices' problem solving performance (Goldsmith & Johnson, 1990).

Structural learning measures have unique measurement advantages. Conventional assessment procedures such as multiple choice tests do not measure knowledge structures (Goldsmith & Johnson, 1990). Multiple choice or completion test items measure the trainee's attainment of individual concepts or propositions. However, such tests do not measure the strength and type of relationship between concepts and propositions. For example, conventional tests can determine if a learner has mastered concepts such as positive reinforcement, punishment, or shaping. Structural assessments, on the other hand, can identify the trainee's associations between these concepts; whether shaping is very similar to, leads to, hinders, or is a type of positive reinforcement.

Purpose: The Need for Reliability Studies of Structural Measures

Empirical research has supported the predictive validity of structural representations such as Pathfinder, card sorts, and multi-dimensional scaling (Gonzalvo, Canas, & Bajo, 1994; Jonassen et al., 1993). That is, structural learning measures have demonstrably predicted changes in memory retrieval., changes in learning, and novice-expert performance shifts (Gonzalvo et al., 1994).

Structural measures are increasingly prevalent as a learning measure. For example, Pathfinder measures are now widely accepted and used as a measure of mental models achievement. Within the last two years The Journal of Educational Psychology alone has published four Pathfinder-related studies of structural learning (e.g., Gonzalvo et al., 1994). An alternative structural learning measure, sorting concepts printed on cards, is a time-honored knowledge assessment technique in psychology that has been used in a number of structural learning studies (e.g., Hirschman & Wallendorf, 1982).

Even though some degree of relationship between these structural learning measures and outcomes have been demonstrated, it is unclear how much the validity of these measures may have been constrained by their unreliability. Surprisingly, there have been few reliability studies of any type on any of the structural learning measures. Ruiz-Primo and Shavelson (1995) note that their survey of concept mapping techniques (card sorts, semantic networks, and Pathfinder measures) turned up only one test reliability study, a study by Lay-Dopyer and Beyerback (1983). Ruiz-Primo and Shavelson concluded that "...reliability of concept map scores is an important issue that must be addressed before they are reported to teachers, students, the public or policy makers." (1995, p. 29).

The few structural learning reliability studies published indicate that test reliability may be a problem. Goldsmith and Johnson (1990) note that their preliminary study of Pathfinder reliability indicated a correlation measure of .60 between subjects' item choices, not a high test reliability index. However, they indicate that the statistic may reflect a lack of test-retest stability due to learning (p. 252). The same-item judgments appear to be taken about six weeks apart with intervening instruction on the topic, contaminating the estimate of reliability.

In summary, structural learning has gained popularity as an important learning outcome, an outgrowth of cognitive research and theory. Structural learning measures are widely used as indicators of structural learning, but their reliability has not been established.

Study Synopsis

This study investigated one type of reliability, test-retest stability, of two widely-used structural learning measures, card sorts of concepts and the Pathfinder program called Knot-Mac or PCKnot. The Knot program, available for Macintosh and Windows environments, was selected because it is an increasingly popular structural learning measure (Schvaneveldt, 1990), used in numerous recent studies of mental model learning; card sorts were investigated because of their traditional and continuing widespread use.

Research Hypothesis

In addition to assessing these structural learning measures' test-retest stability, this study investigated the relationship between subject matter expertise and stability. More specifically, we hypothesized that:

> *Stability will increase with student learning.* Students with little or no knowledge of course concepts will not tap any long-term memory of them, relying on guesses to make similarity judgments or place concepts into groups. As students learn more about the concepts in question their knowledge structures will become more stable, reflecting the permanence of expert knowledge structures (Shavelson 1972). As students learn about the course concepts, their Pathfinder test-retest ratings and their placements of the concepts into categories will reflect long-term memory that is immune to loss in the distracter task.

No direct comparison of the card sort and Pathfinder measures was attempted, as test-retest stability are assessed quite differently for these two measures. Whereas Pathfinder elicits similarity judgments on which test-retest correlations can be calculated, card sorts elicit placement judgments only. Stability for the card sort results is assessed as the agreement in the categories produced between two different sorts. The results from each of these efforts are described separately, with the next section describing the assessment of Pathfinder test-retest reliability.

I. PATHFINDER RELIABILITY STUDY

Reliability measures

Pathfinder reliability will be determined by comparing two different retest data sources, item choice correlations and network similarity.

Correlation measures

Using a test-retest method, reliability was primarily measured by the correlation of a student's concept rating response at time tl to their response at time t2 to the same question.

Item choice correlations reflect the acknowledged test-retest reliability procedure, where one correlates scores of two separate administrations of the same test (Carmines & Zeller, 1979). To assess Pathfinder reliability, this study correlated a subject's similarity ratings of two concepts at time tl to the subject's rating of them at time t2. The two administrations were given within a short time period to minimize reliability threats of student concept change (Carmines & Zeller, 1979, p. 39). A series of intervening distracter tasks minimized memory threats (Ibid.) by clearing working memory of students' recent concept rating judgments.

Similarity measures

Similarity comparisons reflect the correspondence between students' conceptual structures at time tl to time t2. In Pathfinder networks, structural similarity is measured by neighborhood coherence and path distance (Goldsmith & Davenport, 1990). Similarity is the congruence between two networks' conceptual neighborhoods, as determined by the ratio of their intersection to their union. Goldsmith & Johnson (1990) note that similarity measures are the best measures of Pathfinder stability, since they are the best predictors of classroom performance (p.253).

Method

Subjects

The subjects were 18 students enrolled in a graduate educational psychology course at a mid-size southeastern university. The course is an introductory educational psychology course, and assumes no student entry knowledge of psychology. Based on their vitae, 12 students were teacher education students, with four instructional design students and two military training students.

Materials

The pathfinder software used in this program was the Knot pathfinder program (Schvaneveldt, 1990). The study used both the Macintosh (Knot-Mac) and Windows (PCKnot) versions of the program. The program has students to rate the similarity of pairs of concepts. Ten students used the Macintosh version of the program and eight used the Windows version, but all students used the same rating index.

The Knot program was set with *q* = (n-1) and *r* = infinity, the settings used in the majority of Knot-based Pathfinder studies (Johnson, Goldsmith, & Teague, 1995).

Pathfinder administration procedure

First administration - Week 1

1. The first night of class, students met in class, introduced themselves to the class. They then completed a self-assessment that asked them to rate their knowledge of course concepts, the same concepts that would be used in the Pathfinder exercise (Appendix A).

2. After completing the self assessment, students moved to a computer lab. The purpose and procedures of using Pathfinder Knot software was explained to the group. Students were told they were to rate some course concepts as part of an unguarded learning experience. Using an overhead transparency and a white board, the instructor then led the class through several practice concept rating examples involving pets.

3. Using the Knot program, students rated the similarity between 12 educational psychology concepts covered in the course. Students rated the concepts on a 9-point scale, with 9 being most similar and 1 least similar. If the student did not see a relationship between the concepts (because they did not know one or both concepts), they were to mark a midpoint "5" on the scale.

4. The first student to complete the rating task finished in approximately 11 minutes, the last in approximately 22. When finished, students were asked to preview their syllabus for an impending discussion of class management details.

5. After the last student finished, the students performed a distracter task to empty their working memory of their Knot responses. The class reviewed the course syllabus for approximately 20 minutes, sifting at their computers. Class management topics were discussed such as tests, schedules, assignments, and textbooks. Content discussion was avoided, to prevent learning contamination of the upcoming second Knot exercise that night.

6. Students then went on a 10-minute break. When they returned they completed the second Knot rating task, using the same concepts and rating scale. When students were finished they waited at their seats until other class members were done.

Second administration - Week 4

The Week 4 administration followed the sequence of the Week 1 administration. The differences were: 1) no orienting exercise was given, since students seemed familiar with the approach, 2) the distracter task was a 30-minute orientation to the World Wide Web.

Third administration - Week 9

The Week 9 administration followed the sequence of the Week 4 administration. No Pathfinder orientation was needed or given, and the distracter task was a 30-minute (app.) Web search.

Results

Item Choice Correlations

Item-choice correlation data is the most direct and accepted reliability estimate. Table 1 indicates the mean correlation between Pathfinder ratings for the class at Weeks 1, 4, and 9. Only sixteen of eighteen subjects' data are used.

Table 1.
Reliability Estimates: Correlation Between Concept Pair Ratings

	Group Mean Correlation	Standard Deviation	n
Pathfinder 1 - 2 (1st class week)	.506	.134	16
Pathfinder 3 - 4 (4th class week)	.627	.191	16
Pathfinder 5 - 6 (9th class week)	.609	.244	16

One subject missed a mid course Pathfinder exercise and another somehow corrupted their data file, preventing course-wide comparisons of their performance.

The correlation ratings showed a considerable increase between the first and fourth week, with a moderate and nonsignificant drop between the fourth and ninth week (see Table 4). These correlations are very close to the .60 figure mentioned in Goldsmith and Johnson's (1990) Pathfinder learning study. In all cases, the reliability estimates are considerably below Carmines' and Zellars' (1979) requirement of a .80 test-retest correlation figure.

Structural Similarity

Pathfinder reliability was also investigated using structural learning as a secondary, weaker, measure. Structural similarity is determined by the proportion of shared links between students' first and second Pathfinder ratings to the overall number of links in both. This measure is used as a secondary, indirect measure of reliability, since it does not follow the reliability paradigm of directly comparing individual student item choices. Students' individual item ratings could vary while the network similarity rating remains the same. For example, a student choosing a "6" rating for a given concept pair and then an "8" rating for the same pair next time may have the same similarity rating between networks, because the overall structural relationship is maintained. That is, other item choices will maintain the original structure.

Table 2 indicates the similarity of subjects' networks bit Weeks 1, 4, and 9. Similarity ratings were initially low at the outset of the course, with moderately acceptable ratings by the end of it. The similarity of subjects' networks had its largest (and significant) increase between the first and fourth weeks, following the correlation ratings' pattern. The ninth week scores showed a moderate but nonsignificant increase from fourth week scores (Table 4).

Table 2.
Reliability Estimates: Similarity Between
1st, 4th, and 9th Week Student Networks

	Mean Similarity	Standard Deviation	n
Pathfinder 1 - 2 (1st class week)	.388	.132	16
Pathfinder 3 - 4 (4th class week)	.479	.126	16
Pathfinder 5 - 6 (9th class week)	.518	.124	16

Prior Knowledge Effects

Entry level knowledge might variably effect students' first-week reliability indices (network correlation and similarity scores). To measure the effects of students' prior knowledge upon reliability scores, their entry knowledge scores were regressed upon their initial reliability and similarity scores (Table 3).

Table 3.
Correlation Between Prior Knowledge And Pathfinder Correlation And Similarity Scores

	r^2	t	p
Prior Knowledge regressed upon Correlation 1 - 2	3.6%	-.75	.465
Prior Knowledge regressed upon Similarity 1 - 2	1.6%	.497	.626

n = 17

Students' entry knowledge was obtained from a self-assessment survey that students completed the first night of class. The self-assessment measure had students rank their initial familiarity with the course concepts used in the Knot rating task. Pretest ratings could not be used, since pretests assess the correctness of knowledge more than its long-term stability. That is, a student may have an incorrect (by pretest) but stable concept of a topic. The stability of subjects' concepts more directly corresponds to their familiarity with them.

The initial reliability scores were derived from students' first-night Pathfinder networks. The regressions indicated that there was no relation between prior knowledge and either reliability score. Consequently, prior knowledge scores were dropped from further statistical calculations.

Affects of Course Learning Upon Reliability Estimates

The initial hypothesis of this study was that reliability scores will increase with student learning and the subsequent stabilization of course concepts in long-term memory. To investigate this, we compared students' first, fourth, and ninth week reliability scores, as measured by correlation and similarity. Table 4 summarizes these results. Reliability estimates markedly increased with student learning from the first to fourth weeks, as measured by correlation and similarity consistency. The learning interval was only 3 weeks, but noticeable changes ensued, a 23% increase in the mean correlation for both measures.

Table 4.
Changes in Reliability Estimates:
Correlation and Similarity Differences Between
1st, 4th and 9th Week Group Means

	1st to 4th Week	4th to 9th Week	1st to 9th Week
Mean correlation between item ratings (n=16)	.506 to .627**	.627 to .609	.506 to .609*
Similarity between structures (n = 16)	.388 to .479*	.479 to .518	.388 to .518**

One-tailed t-test for dependent (paired) means. *p <.05 **p < .005

Fourth-to-ninth week reliability estimates showed no significant increases, even though the learning period was longer than the first-to-fourth week period. Part of the reason for this stabilization of Pathfinder reliability may be that nine of the eleven concepts rated by Pathfinder (Appendix A) were covered by students by the fourth week of class, since their fourth-week readings had an overview of forthcoming course topics.

Discussion

This Pathfinder study has produced some intriguing results, indicating that Pathfinder reliability warrants further exploration for two reasons: 1) to evaluate its general feasibility as a reliable measurement instrument, and 2) to clarify Pathfinder content and implementation issues that will improve its reliability for individual administrations of it. Some specific conclusions follow:

Pathfinder reliability is questionable. None of the correlation or similarity ratings indicated a high consistency between student choices, regardless of the time period administered. Future reliability studies are warranted. At present, multiple Pathfinder reliability indices should be used in any learning experiment that utilizes this measure (e.g. correlation and similarity indices). Where possible, a secondary structural learning measure should also be used with Pathfinder ratings (e.g., semantic networks, concept maps, etc.) .

Future studies should further investigate Pathfinder reliability with abstract and concrete concept sets, and with concept sets that have high and low degrees of pre-experimental learner familiarity.

Pathfinder reliability may increase with student learning. Correlation and similarity ratings showed their greatest increases during the first-to-fourth week learning period. These data supported the hypothesis that learning may increase the stability (reliability) of the Pathfinder measure, at least during the first weeks of learning. However, measures of subsequent (ninth week) structural learning showed no increase in reliability by either correlation or similarity measures. This stability may be due to the early coverage of Pathfinder concepts in the course, or due to students' acquisition of a general schema by the fourth week.

Future studies should replicate this reliability study with concept sets that include more end-of-course topics, to determine if reliability indices increase with course learning.

Multiple Pathfinder administrations may increase reliability. Students who use Pathfinder may need to cognitively acclimate themselves to the Pathfinder interface and its conceptual task of pairing isolated concepts. This may be true even when students first practice with an unrelated concept set.

Future studies should investigate the correlation and similarity between student networks after students have rated: (1) a "first run" of the concept set of interest, or (2) a different concept set of the same domain and level of abstraction.

Pathfinder measures such as Knot-Mac and PCKnot may benefit from the addition of a "don't know" option in similarity ratings. On standard test of student' learning, student ignorance is not a reliability issue. A student who does not know a concept is expected to miss questions about it. A test-retest reliability measure of these questions is not diminished if the student chooses a different type of wrong answer the second time.

A future study would use a Pathfinder interface that has a "don't know option" built into the interface, and compare reliability and predictive validity of the new interface with standard Pathfinder similarity ratings provided by programs such as Knot-Mac or PCKnot.

II. CARD SORT STABILITY STUDY

Stability assessment

A number of statistics are available to assess the agreement between two groupings, based on pairwise classification of the items in the two solutions. One such statistic is the Fowlkes and Mallows (1983). If A indicates the number of pairs of tasks grouped in both solutions, and B and C indicate frequencies of disagreement in which pairs are grouped by one method, but not the other, the Fowlkes and Mallows (F&M) is computed as follows:

$$\text{F\&M} = \frac{A}{\sqrt{(A + B) * (A + C)}}$$

The F&M has an upper bound of 1.00 when the two solutions agree perfectly, and a lower bound of 0.0 when A is zero. Additionally, it is undefined when cells A, B, and C are all equal to zero, but this would occur only when the number of groups equals the number of cases for both solutions (i.e., no grouping had occurred).

To evaluate changes in the stability of the students' groupings during the semester, F&M agreement statistics were computed for each of the student's pairs of sorts. These F&M statistics were averaged for each of the 3 assessment periods during the semester, and changes in stability were tested with a one-tailed t-test for paired means.

The relationship between these stability measures and student learning was examined in several ways. The first approach was to assess the consistency between the students' groupings during the semester and groupings formed by the instructor. The F&M, like other agreement statistics of this type, however, is affected by differences in the size of groups being compared. With random sortings only, two solutions with larger groups will yield a higher F&M than two solutions with smaller groups. To help control for this effect due to the specificity of the solution, the instructor, prior to the course, formed a hierarchical grouping of the concepts, so that the preferred solution at any level of specificity could be identified. F&M agreement statistics were then computed between each of the student's sorts and the instructor's sort with the same number of groups. These F&M statistics were averaged for each of the 3 assessment periods during the semester, and changes in consistency with the instructor were tested with a one-tailed t-test for paired means.

Pearson product moment correlations between several other course performance measures and average F&M statistics between student sorts (stability) and between student and instructor sorts (consistency) were also computed. These performance measures included the student's quiz average, score on the final exam, and total course percentage. In addition, students rated their familiarity with each of the concepts at the beginning and end of the course on a 5-point scale, from "very unfamiliar" to "very familiar". Correlations between stability, consistency, and these self-reported familiarity ratings were also computed.

Finally, the nature of structural learning for a group of subjects has often been characterized by cluster analyzing card sort data (e.g., Schoenfeld & Herrmann, 1982). For this study, the proportion of times the concepts were grouped together at each assessment period was used as a measure of similarity in a hierarchical cluster analysis using average linkage between groups. The cluster diagrams based on the data from the first and final pairs of card sorts are reported.

Method

Subjects

The subjects were 10 students enrolled in an undergraduate class on learning and memory in a large Midwestern university. The course was part of the university's night school, and all of the students were working adults. The class was composed of 3 freshman, 1 sophomore, 3 juniors, and 3 seniors, and all were psychology majors. None, however, had previous course work in learning and memory.

Procedure

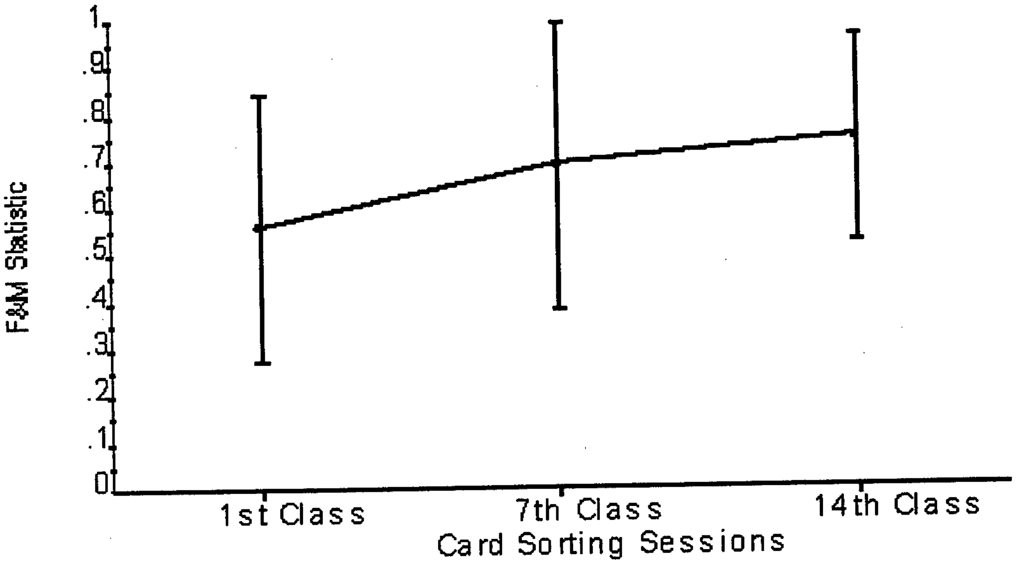
Perhaps one of the most straightforward ways to assess knowledge structure is simply to have people sort concepts printed on cards into categories. The groups that result constitute the important organizing structures for the individual. This general card-sorting approach has a long history in psychology, in applications ranging from industrial psychology where it is used for training needs assessment (Goldstein, 1993), to personality, attitude, or preference assessment, originating with Stephenson's (1953) Q sort method. More germane to the current study, card sort measures have also been used to assess structural learning (e.g., Hirschman & Wallendorf, 1982; Schoenfeld & Herrmann, 1982).

Twelve concepts from the course were selected for the card sorting exercise (contained in Appendix B). For each sort, the subjects were asked to sort the concepts into groups of similar items. They were not told how many groups to form, but rather that they should use their judgment in determining the best set of groups. In the first week of the course, the subjects performed the first two sorts. Approximately 30 minutes passed between these tasks, during which the course syllabus was reviewed. The subjects performed a second and third set of two groupings during the seventh and fourteenth class meetings, with the sorts again being separated by 30 minutes of other activities. In no case did the filler activities involve any discussion of the concepts to be grouped. Thus, 6 card sorts were performed in all, with 2 at the beginning, 2 at mid-semester, and 2 at the end of the course.

Results

The stability of the students' groupings tended to increase, with the t-test between the first and second ($t = 1.52$, $p < 0.10$) and between the first and last card sorts ($t = 1.41$, $p < 0.\ 10$). These means, with one-standard deviation bars, are illustrated in Figure 1. This tendency toward an increase in stability may be produced by learning about the concepts; that is, the students may be able to tap long-term memory structures regarding the items and how they relate to other concepts on the list during the latter card sorts, resulting in greater test-retest stability of the groupings.

Figure 1.
Changes in the stability of students' groupings during the semester.

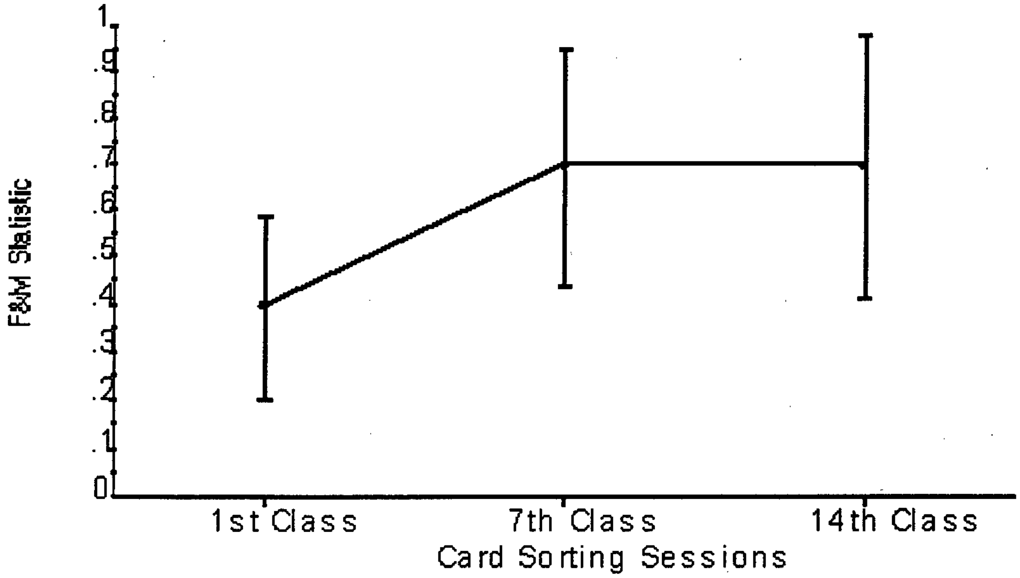


It is also possible, of course, that the groupings became more stable simply from the repetition of the card sorting exercise. In this case, the groupings may be idiosyncratic and largely unrelated to other measures of learning in the course. The subsequent analyses address this possibility.

One of the crucial pieces of information indicating that the observed increase in stability results from learning relationships among the concepts is change in consistency between the instructor's initial grouping and the students' groupings over the semester. Consistency was found to increase during the semester, with the t-test between the first and second ($t = 4.51$, $p < 0.01$) and between the first and last sessions ($t = 2.75$, $p < 0.01$) reaching statistical significance. Thus, not only were the sortings tending toward greater stability, they were also converging on the instructor's initial groupings. The means of the F&M statistics measuring the consistency between student and instructor groupings, with one standard deviation bars, are shown in Figure 2.

It is noteworthy that the change in consistency, and to a lesser extent, the change in stability, appears to disappear after mid-semester. A review of the list of concepts revealed that all but one of the items had been covered by midsemester. Although this bias toward concepts covered early in the course was unintentional, it provides a likely explanation for the lack of change later in the semester.

Figure 2.
Changes in the consistency between the instructor's initial grouping and the students' groupings during the semester.



If learning is producing the trend toward increased test-retest stability in the card sort data, stability should also be correlated with other, more traditional measures of learning. Correlations between some of these measures and both stability between a student's sorts and consistency between a student's and the instructor's sorts are shown in Table 5. In general, both stability and consistency correlated with course learning measures such as quiz and final scores and the total course percentage.

Unexpectedly, the correlations between these traditional learning measures and stability at the end of the semester dropped. This reversal may have occurred, however, because the student with the highest average reported changing her approach to organizing the concepts during the final card sorting exercise. The implications of this shift in strategy are further discussed in the conclusions section.

Correlations between self-reported familiarity and both stability and consistency are also reported in Table 5. Interestingly, initial familiarity correlated poorly, and predominantly negatively, with both stability and consistency throughout the semester. Apparently, initial student beliefs about

Table 5.
Correlations between stability, consistency, and several other course performance measures

Stability at	Quiz	Final	Total	Initial Familiarity	Final Familiarity
1st Class	.73*	.62	.67*	-.33	.24
7th Class	.75*	.68*	.66*	-.48	.82*
14th Class	.32	.20	.18	-.38	.48
Consistency at					
1st Class	.62	.77*	.71*	.16	.28
7th Class	.78*	.72*	.72*	-.47	.65*
14th Class	.66*	.56	.56	-.47	.64*

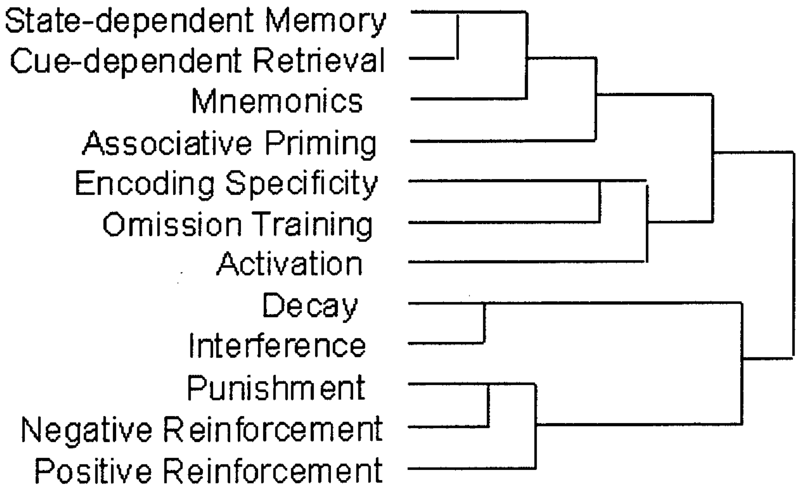
* p < .05

how familiar the concepts were to them were not strongly related to how stable their groupings were over the 30 minute test period or how consistent they were with the instructor's groupings. Final ratings, however, were positive and mostly significant. This result may indicate a change in their perceptions about how well they understand the concepts, or it may merely reflect a rating that is more consistent with the grades and other feedback they had been receiving.

When the structure of concepts for a group is expected to change, cluster analysis based on card sort data has been used to characterize these patterns at different points in time. Schoenfeld and Herrmann (1982), for example, found that the initial clustering of mathematics problems based on card sorts tended to reflect surface characteristics. After a class covering these problems, however, the cluster diagram reflected a structure based on the approach to solving them.

The cluster diagram based on the initial card sortings is illustrated in Figure 3. The structure, in part, reflects the students' previous knowledge of learning and memory. For example, the grouping of punishment, negative, and positive reinforcement fairly early in the cluster diagram reflects a set of concepts from instrumental conditioning. Similarly, decay and interference are major viewpoints of forgetting.

Figure 3.
Cluster diagram based on card sorts from the beginning of the semester.



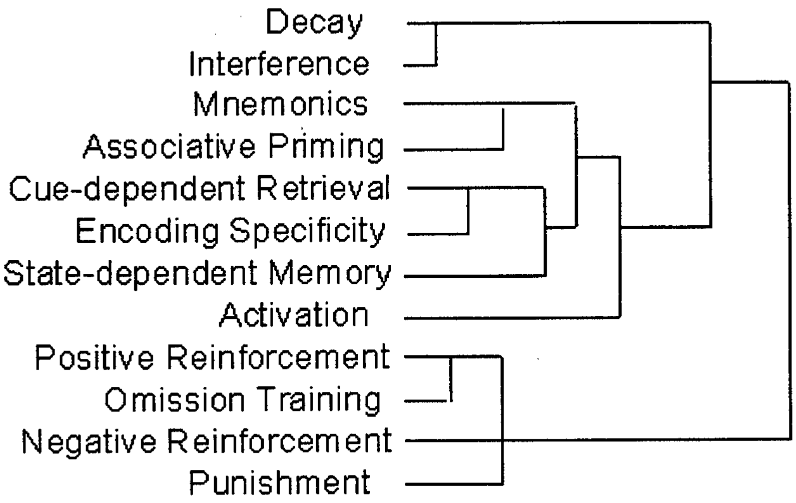
But also reflected in Figure 3 are some apparent misunderstandings of State-dependent Memory, Cue-dependent Retrieval, Mnemonics, Associative Priming, Encoding Specificity, Omission Training, Activation, Decay, Interference, Punishment, Negative Reinforcement, and Positive Reinforcement. For example, the fourth instrumental conditioning concept, Omission Training, is not initially grouped with the other three. Decay and Interference are eventually grouped with the concepts from instrumental conditioning, although the remaining concepts largely reflect ideas associated with the cue-dependent retrieval theory of forgetting. And finally, the close association of Cue-dependent Retrieval and State-dependent Memory may, in part, reflect the similarly of the terminology, rather than the similarity of the concepts.

The final card sort data were also used to cluster the concepts, and the resulting diagram is shown in Figure 4. Although the diagram is not totally parallel to the concept structure formed by the instructor before the class began, most of the differences are relatively minor.

Discussion

Overall, the results from this study suggested that the stability of card sorts as an indicant of structural learning increased as a result of acquiring knowledge about them. With a more complete understanding of the concepts at the end of the course, the students used long-term memory structures to form two similar card sorts during the test-retest session.

Figure 4.
Cluster diagram based on card sorts from the end of the semester



Although the change in stability was significant only at the alpha . 10 level, this result is perhaps to be expected, based on the students' previous exposure to the concepts. According to the initial cluster analysis, the students already understood the relationships among several of the concepts. Instability, then, could only be reflected by placement of the unfamiliar items. This results contrasted substantially with the findings on consistency, where the sessions at the middle and end of the semester produced results that were statistically different from those from the first of the semester. Over the course of the semester, the groupings became somewhat more stable, and much more consistent with the instructor's classifications of the concepts. And consistency, in particular, correlated well with other, more traditional measures of class performance.

Perhaps one of the more important findings, however, was the lack of correlation between these traditional performance measures and the stability of the card sorts at the end of the semester. Although only anecdotal evidence is available, the student with the highest percentage reported changing her organizational strategy between the test and the retest during the last experimental session. Additionally, she reported altering her strategy as a result of reflecting on the previous card sort; that is, the measurement process affected the quantity being measured. This change would reduce stability, and so, lowered the correlation with traditional performance measures. In fact, the F&M for her last two card sorts was only 0.478. If this single data point is omitted, the remaining stability estimates at the end of the semester correlated between .55 and .62 with the other course performance measures.

Similar effects have been observed elsewhere. In a card sort study by Hirschman and Wallendorf (1982), only 50% of their subjects reported using the same organizing scheme in successive card sort tasks. If the act of performing a card sort stimulates reflection, and in some cases, change in organizational strategies, the measuring method may itself produce some instability, and require more repetition to produce a stable assessment.

CONCLUSIONS

Pathfinder networks continue to grow in popularity as a measure of structural learning, while card sorting continues to see widespread use. This study supports the conclusion that these methods can produce stable, coherent measures of structural learning, if properly applied. It does not, however, imply that these results will always follow. The stability of card sorts and Pathfinder networks as measures of structural learning will most likely depend upon a host of factors, such as the content of the concept list, the subjects' prior knowledge about the concepts, the method of administration, and so on. Additional work is planned to help develop test construction and administration procedures that improve the reliability of the measure for each experiment in which it is employed.

Apart from such guidance, however, one should seek to establish the stability of any specific application of a structural learning measure.

Unfortunately, this study implies that obtaining this information may be problematic. In measuring structural learning, learning may be affected, so that the act of assessing stability may, in some cases, produce instability. Additional work should be devoted to developing a more standardized approach to assessing the stability of structural learning measures before they are widely used as learning criteria.

REFERENCES

Anderson, J. R. (1995). *Learning and memory.* New York: John Wiley & Sons.

Carmines, E.G. & Zeller, R.A. (1979). Reliability and validity assessment. Beverly Hills, CA: Sage Publications.

Fowlkes, E. B. & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings, (with comments and rejoinder). *Journal of the American Statistical Association, 78,* 553-584.

Goldsmith, T. E. & Johnson, P. (1990). A structural assessment of classroom learning. In R. Schvaneveldt (ed.) *Pathfinder associative networks: studies in knowledge organization.* (pp. 241-254). Norwood, New Jersey: Ablex.

Goldsmith, T. E. & Davenport, D. M. (1990). Assessing structural similarity of graphs . In R. Schvaneveldt (ed.) *Pathfinder associative networks: studies in knowledge organization.* (pp. 241-25). Norwood, New Jersey: Ablex.

Goldstein, I. L. (1993). *Training in organizations* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing.

Gonzalvo, P., Canas, J.J., & Bajo, M. (1994). Structural representations in knowledge acquisition. Journal of Educational Psychology, 86, 601-616.

Hirshman, E. C. & Walendorf, M.R. (1982). Free response and card-sort techniques for assessing cognitive content: Two studies concerning their stability, validity, and utility. *Perceptual and Motor Skills,* 54, 1095-1110.

Johnson, P. Goldsmith, T., & Teague, K. W. (1995). Similarity, structure, and knowledge: A representational approach to assessment. In P. Nichols, S. Chipman & R. Brennan (Eds.) *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

Jonassen, D., Beissner, K. & Yacci, M. (1993). *Structural knowledge: techniques for representing, conveying and acquiring structural knowledge.* Hillsdale, New Jersey: Lawrence Erlbaum.

Jonassen, D. & Tessmer, M. (in press). An outcomes-based taxonomy for design, evaluation, and research of instructional systems. *Training Research Journal,* in press.

Kraiger, K., Ford, J., K., & Salas, E. (1993) Integration of cognitive, behavioral, and affective theories of learning into new methods of training evaluation. Journal of Applied Psychology, 78, 311-328 (Monograph).

Lay-Dopyera, M., & Beyerbach, B. (1983) Concept mapping for individual assessment (ERIC Document Reproduction Service No. ED 229 399)

Rumelhart, D. E. & & Norman, D. A. (1981). Accretion, tuning and restructuring: three modes of learning. In J. W. Cotton and R. Klatzky (Eds.) *Semantic factors in cognition* (pp. 37-54). Hillsdale, NJ: Erlbaum.

Ruiz-Primo, M. A., & Shavelson, R. J. (1995, April). *Concept maps as potential alternative assessments in science*. Paper presented at the annul convention of the American Educational Research Association, San Francisco, CA.

Schoenfeld, A.H. & Herrmann, D. J. (1982). Problem perception and., knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology. Learning, Memory and Cognition, 8,* 484 - 494.

Schvaneveldt, R. (1990). Ed - *Pathfinder associative networks: studies in knowledge organization.* Norwood, NJ: Ablex.

Shavelson, R. J. (1972). Some aspects of the correspondence between content structure and cognitive structure in physics instruction. *Journal of Educational Psychology,* 63, 225-234.

Stephenson, W. (1953). *The study of behavior.* Chicago: University of Chicago Press.

Appendix A
Course Concepts Used in Pathfinder Ratings Study

positive reinforcement
negative reinforcement
punishment
modeling
cognitivism
long term memory
short term memory
learning strategies
motivation
social learning theory
behaviorism

Appendix B
Course Concepts Used in Card Sort Study

activation
state-dependent memory
mnemonics
interference
omission training
negative reinforcement
associative priming
punishment
positive reinforcement
decay
cue-dependent retrieval
encoding specificity